Deep learning-based artificial intelligence model to assist thyroid nodule diagnosis and management: a multicentre diagnostic study

Sui Peng, Yihao Liu*, Weiming Lv*, Longzhong Liu*, Qian Zhou*, Hong Yang, Jie Ren, Guangjian Liu, Xiaodong Wang, Xuehua Zhang, Qiang Du, Fangxing Nie, Gao Huang, Yuchen Guo, Jie Li, Jinyu Liang, Hangtong Hu, Han Xiao, Zelong Liu, Fenghua Lai, Qiuyi Zheng, Haibo Wang, Yanbing Li, Erik K Alexander, Wei Wang, Haipeng Xiao*

Summary

Background Strategies for integrating artificial intelligence (AI) into thyroid nodule management require additional development and testing. We developed a deep-learning AI model (ThyNet) to differentiate between malignant tumours and benign thyroid nodules and aimed to investigate how ThyNet could help radiologists improve diagnostic performance and avoid unnecessary fine needle aspiration.

Methods ThyNet was developed and trained on 18 049 images of 8339 patients (training set) from two hospitals (the First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China, and Sun Yat-sen University Cancer Center, Guangzhou, China) and tested on 4305 images of 2775 patients (total test set) from seven hospitals (the First Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou, China; the Sixth Affiliated Hospital of Sun Yat-sen University, Guangzhou, China; the Guangzhou Army General Hospital, Guangzhou, China; the Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China; the First Affiliated Hospital of Sun Yat-sen University; Sun Yat-sen University Cancer Center; and the First Affiliated Hospital of Guangxi Medical University, Nanning, China) in three stages. All nodules in the training and total test set were pathologically confirmed. The diagnostic performance of ThyNet was first compared with 12 radiologists (test set A); a ThyNet-assisted strategy, in which ThyNet assisted diagnoses made by radiologists, was developed to improve diagnostic performance of radiologists using images (test set B); the ThyNet assisted strategy was then tested in a real-world clinical setting (using images and videos; test set C). In a simulated scenario, the number of unnecessary fine needle aspirations avoided by ThyNet-assisted strategy was calculated.

Findings The area under the receiver operating characteristic curve (AUROC) for accurate diagnosis of ThyNet (0·922 [95% CI 0·910–0·934]) was significantly higher than that of the radiologists (0·839 [0·834–0·844]; p<0·0001). Furthermore, ThyNet-assisted strategy improved the pooled AUROC of the radiologists from 0·837 (0·832–0·842) when diagnosing without ThyNet to 0·875 (0·871–0·880; p<0·0001) with ThyNet for reviewing images, and from 0·862 (0·851–0·872) to 0·873 (0·863–0·883; p<0·0001) in the clinical test, which used images and videos. In the simulated scenario, the number of fine needle aspirations decreased from 61·9% to 35·2% using the ThyNet-assisted strategy, while missed malignancy decreased from 18·9% to 17·0%.

Interpretation The ThyNet-assisted strategy can significantly improve the diagnostic performance of radiologists and help reduce unnecessary fine needle aspirations for thyroid nodules.

Funding National Natural Science Foundation of China and Guangzhou Science and Technology Project.

Introduction

Thyroid nodules are found in up to 68% of asymptomatic adults in the general population.[1] Approximately 7–15% of thyroid nodules are thyroid cancer, which is the most rapidly increasing malignancy in all populations.[2] The large number of thyroid nodules, with only a fraction being cancerous, calls for a reliable method to accurately differentiate malignant from benign nodules.

Routine decision making for patients with thyroid nodules depends on ultrasound or invasive fine needle aspiration.[2] However, the assessment of ultrasound features is time consuming, subjective, and often dependent on a radiologist's experience and the available ultrasound devices.[3] Ultrasound conclusions are often inconsistent and even with fine needle aspirations 15–30% of the samples still yield indeterminate cytological findings.[4] Additional robust methods are needed to improve diagnosis and fine needle aspiration strategies to adapt to the exponential growth of patient needs and burden on medical services.

Artificial intelligence (AI) has been reported to meet or exceed human experts in medical imaging.[5–8] A few

Lancet Digit Health 2021; 3: e250–59

*Contributed equally to this work

Clinical Trials Unit (Prof S Peng PhD, Y Liu MD, Q Zhou MS, Prof H Wang MPH), **Department of Endocrinology** (Y Liu, F Lai MM, Q Zheng MD, Prof Y Li PhD, Prof H Xiao PhD), **Department of Medical Ultrasonics, Institute of Diagnostic and Interventional Ultrasound** (Y Liu, J Liang PhD, H Hu MD, Han Xiao MD, Z Liu MD, Prof W Wang), **and Department of Breast and Thyroid Surgery** (Prof W Lv PhD, J Li PhD), **The First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China; Department of Ultrasound, Sun Yat-sen University Cancer Center, State Key Laboratory of Oncology in South China, Guangzhou, China** (L Liu PhD); **Department of Medical Ultrasound, the First Affiliated Hospital of Guangxi Medical University, Nanning, China** (Prof H Yang PhD); **Department of Medical Ultrasonics, the Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China** (Prof J Ren PhD); **Department of Medical Ultrasonics, the Sixth Affiliated Hospital of Sun Yat-sen University, Guangzhou, China** (G Liu PhD); **Department of Medical Ultrasonics, the First Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou, China** (X Wang MD); **Department of Ultrasound, the Guangzhou Army General Hospital, Guangzhou, China** (X Zhang MD); **Xiaobaishiji, Beijing, China** (Q Du ME, F Nie ME, G Huang DE); **Institute for Brain and Cognitive Sciences, Tsinghua University, Beijing, China** (Y Guo ME); **Thyroid Section,**

Brigham & Women's Hospital, Harvard Medical School, Boston, MA, USA (Prof E K Alexander MD)

Correspondence to:
Prof Haipeng Xiao, Department of Endocrinology, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou 510080, China
xiaohp@mail.sysu.edu.cn

or

Prof Wei Wang, Department of Medical Ultrasonics, Institute of Diagnostic and Interventional Ultrasound, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou 510080, China
wangw73@mail.sysu.edu.cn

or

Prof Erik K Alexander, Thyroid Section, Brigham & Women's Hospital, Harvard Medical School, Boston, MA 02115, USA
ekalexander@bwh.harvard.edu

Research in context

Evidence before this study

We searched PubMed from the inception of the database to Sept 20, 2020, for research articles with the search terms "deep learning" OR "machine learning" OR "artificial intelligence" OR "convolutional neural network" AND "thyroid cancer" OR "thyroid nodule" OR "thyroid carcinoma", without language restrictions. We identified 15 studies on the development and validation of artificial intelligence (AI) models in thyroid nodule management. However, these studies compared the performance of radiologists with that of the AI model. We found no publications that specifically reported how diagnostic deep-learning or machine-learning algorithms could assist radiologists performance in thyroid nodule management. The absence of multicentre training cohorts and a small number of ultrasound devices in previous studies restricted their generalisability in clinical practice.

Added value of this study

To our knowledge, this study is the first to develop an AI-assisted strategy for thyroid nodule management. The ThyNet-assisted strategy not only improved the performance of radiologists when reviewing images only, but also when reviewing images and videos in a clinical setting. Of note the combination of the American College of Rheumatology Thyroid Imaging Reporting and Data System classification with AI assistance improved the negative predictive value and positive predictive value of thyroid nodule differentiation, which reduced the number of unnecessary fine needle aspiration.

Implications of all the available evidence

ThyNet-assisted strategy could significantly improve the diagnostic performance of radiologists and help reduce the number of unnecessary fine needle aspirations for thyroid nodules. On the basis of our findings, AI diagnostic programmes should be rolled out to clinical practice of thyroid nodule management.

studies have focused on a comparison of the diagnostic performance of AI with clinicians in thyroid nodule differentiation.[9–11] In our preliminary study, a machine learning system showed a better predictive value for malignant thyroid nodules compared with humans using American College of Rheumatology (ACR) Thyroid Imaging Reporting and Data System (TI-RADS).[7] The introduction of deep learning in thyroid imaging has also achieved a better diagnostic performance than experienced radiologists.[12,13] Previous studies applying deep-learning algorithms have mainly focused on the comparison of radiologists and deep-learning models by reading ultrasound images. However, in a real-world setting, the final diagnosis should still be made by radiologists. Therefore, evaluating the diagnostic improvements provided by the cooperation between radiologists and AI systems is more similar to the clinical setting. Radiologists could improve performance by reading dynamic videos instead of static images only, but whether an AI-assisted model can help radiologists improve diagnostic performance by processing both images and videos should be investigated. Moreover, few studies discussed the influence of AI on fine needle aspiration or thyroidectomy treatment advice given by health-care professionals, leaving this issue still vague.

We developed a deep-learning AI model (ThyNet) to differentiate malignant tumours from benign thyroid nodules. We investigated whether radiologists could improve their diagnostic performance with the assistance of the ThyNet model when reading ultrasound images and videos and explored the potential of the ThyNet-assisted strategy to help radiologists avoid unnecessary fine needle aspirations.

Methods

Study design and datasets

This was a multicentre, diagnostic study that used ultrasound image sets from seven hospitals in China. Patients aged 18 years old or older with thyroid nodules at least 3 mm in diameter identified with ultrasound who had a definitive benign or malignant pathological result (surgical specimen or fine needle aspiration [Bethesda category II or VI]) were eligible for inclusion in the training set and testing sets. The pathological diagnoses were made by two pathologists, one of whom had more than 8 years' experience. All images were intially included, but low-quality ultrasound images, such as severe artifacts (eg, motion artifacts and speed propagation and refraction artifacts) or low image resolution, were excluded after screening.

The images of the training set were collected from the First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China and Sun Yat-sen University Cancer Center, Guangzhou, China (18049 images of 8339 patients). For test set A, 2185 images of 1424 patients with thyroid nodules were enrolled from four independent hospitals (the First Affiliated Hospital of Guangxi Medical University, Nanning, China; the First Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou, China; the Sixth Affiliated Hospital of Sun Yat-sen University, Guangzhou, China; and the Guangzhou Army General Hospital, Guangzhou, China). For test set B, 1754 images of 1048 patients with thyroid nodules were enrolled from the First Affiliated Hospital of Sun Yat-sen University, and the Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China. For test set C, 366 images of 303 patients with thyroid nodules were enrolled from the First Affiliated Hospital of Sun Yat-sen University, Sun Yat-sen University Cancer Center, and the First Affiliated Hospital of Guangxi Medical University.

This study was approved by the Research Ethics Committee of the First Affiliated Hospital of Sun Yat-sen University. Informed consent was waived for retrospectively collected ultrasound images, which were annonymised. Written informed consent was obtained from patients whose ultrasound images and dynamic videos were prospectively collected.

Outcomes

The primary endpoint of our study was the area under the receiver operating characteristic curve (AUROC) of thyroid nodule diagnosis. The secondary endpoints of our study were accuracy, sensitivity, specificity, positive predictive value, and negative predictive value of thyroid nodule diagnosis. The post-hoc analysis included the diagnostic accuracy of ThyNet in different pathological subtypes and ThyNet-assisted fine needle aspiration strategy.

Procedures

For the training set, ultrasound images of consecutive patients with thyroid nodules were retrospectively retrieved from the individual thyroid imaging database at the First Affiliated Hospital of Sun Yat-sen University and Sun Yat-sen University Cancer Center, between Jan 1, 2009, and Nov 30, 2018. A total of 19 312 images from 8339 patients were included in the training set, with 1263 images excluded due to poor image quality.

There was no overlap between patients in the training and test sets and there was no overlap between the three test subset. The test set images for the comparison between ThyNet and radiologists (test set A) and the assessment of the ThyNet-assisted strategy (test set B) were retrospectively obtained from six independent hospitals (the First Affiliated Hospital of Guangxi Medical University, the First Affiliated Hospital of Guangzhou University of Chinese Medicine, the Sixth Affiliated Hospital of Sun Yat-sen University, the Guangzhou Army General Hospital, First Affiliated Hospital of Sun Yat-sen University, and the Third Affiliated Hospital of Sun Yat-sen University) between Jan 1, 2009, and July 30, 2019. In the clinical setting test (test set C), both images and dynamic videos of nodules were prospectively collected from inpatients at the First Affiliated Hospital of Sun Yat-sen University, Sun Yat-sen University Cancer Center, and First Affiliated Hospital of Guangxi Medical University from Oct 1 to Nov 30, 2019 (appendix p 4). A total of 6587 patients in the training set and 1956 patients in the test sets were confirmed as having a definitive benign or malignant pathological result based on a surgical specimen. 1752 patients in the training set and 819 patients in the test sets were confirmed as having a definitive benign or malignant pathological result based on fine needle aspiration (Bethesda category II or VI).

All thyroid ultrasound images extracted from the thyroid imaging database were converted into a JPEG format. Various models of ultrasound equipment produced by 13 different manufacturers (GE Healthcare, Chicago, IL, USA; Philips, Amsterdam, the Netherlands; Siemens, Munich, Germany; Canon, Tokyo, Japan; Samsung, Seoul, South Korea; Esaote, Genoa, Italy; Mindray, Huntingdon, UK; SonoScape, Shenzhen, China; Aloka, Wallingford, CT, USA; BK Medical, Peabody, MA, USA; Supersonic, Aix-en-Provence, France; Vinno, Suzhou, China; and Hitachi, Tokyo, Japan) were used to generate the ultrasound images (appendix p 8). Image quality control was done for the training set and test sets. For the quality control of ultrasound images, all thyroid images were screened and low-quality images containing severe artifacts or significant image resolution reductions were removed. The screening for the images was done by two radiologists (HanX and ZL) who had at least 1 year of ultrasound experience. If there was no consensus regarding nodule location between the image and the pathological report, the image was removed. 2345 images from 1424 patients in test set A for the comparison between ThyNet and radiologists met the criteria, with 160 images excluded. 1896 images from 1048 patients met the inclusion criteria and were used in the assessment of the ThyNet-assisted diagnostic strategy, with 142 images excluded after image quality control (test set B). 401 images from 303 patients in test set C met the inclusion criteria and were used in the assessment of the ThyNet-assisted diagnostic strategy in a real-world setting, with 35 images excluded after image quality control. All data were deidentified (including retrospectivley collected data for the training sets) before the development and evaluation of the model.

The ThyNet deep-learning algorithm was specifically designed to diagnose malignancy from thyroid ultrasound images. It is a combined architecture of three networks: ResNet, ResNeXt, and DenseNet (appendix p 5). ResNet uses residual learning blocks to reduce the effect of gradient vanishing. ResNeXt is a modified version of ResNet, developed by repeating a building block that aggregates a set of transformations with the same topology. ResNeXt additionally introduced the concept of sparsity and group convolution to enhance the ability of the AI to learn the semantic information with less parameters. DenseNet is a new network architecture that connects each layer to every other layer in a feed-forward fashion.[14] DenseNet makes the network deeper but reduces the number of parameters and prevents overfitting. The three branches of networks were trained separately on the same training set and assembled through a majority vote algorithm. To search for the optimal weights for each network branch and get the ensembled output, we used the brute-force search method via cross-test in the training sets. The final weighting ratios are 0·40 for ResNet, 0·35 for ResNeXt, and 0·25 for DenseNet.

See Online for appendix

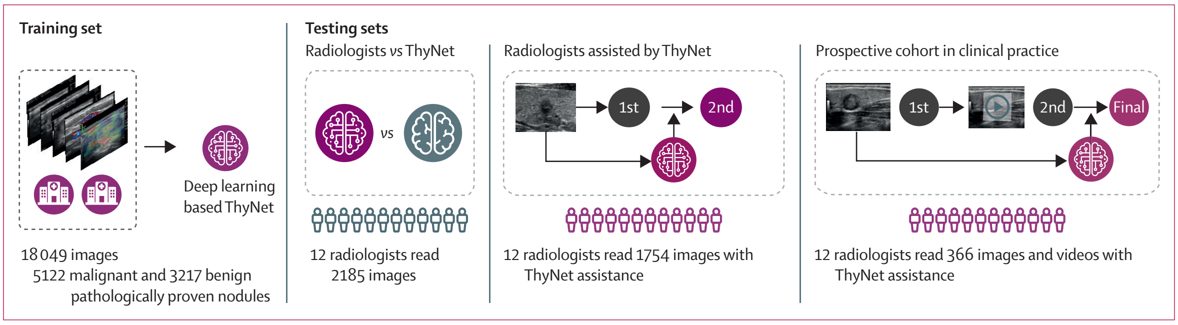
Training set
Deep learning based ThyNet
18 049 images
5122 malignant and 3217 benign pathologically proven nodules
Testing sets
Radiologists vs ThyNet
vs
12 radiologists read 2185 images
Radiologists assisted by ThyNet
1st
2nd
12 radiologists read 1754 images with ThyNet assistance
Prospective cohort in clinical practice
1st
2nd
Final
12 radiologists read 366 images and videos with ThyNet assistance

Figure 1: **Study profile**

Using datasets from two centres, ThyNet was trained to differentiate thyroid nodules. ThyNet was then tested on three datasets with no overlap (test sets A–C). First, diagnostic performance between radiologists and ThyNet based on static images was compared. Second, diagnostic performance of radiologists before (first diagnosis) and after (second diagnosis) the assistance by ThyNet was assessed based on static images. Third, the first diagnosis based on static images and the second diagnosis based on dynamic videos was recorded. Then, with the assistance of ThyNet, the final diagnosis was obtained and compared with the independent diagnoses made by radiologists without ThyNet.

The noise information (eg, paramaters of the ultrasound device), which was distributed mainly in the peripheral areas of the original images, was manually removed by one radiologist (HH). The images were resized to 256×256 pixels before being cropped to 224×224 pixels. Standard image preprocessing (clipping, flipping, and rotating) for deep learning to generate a larger, more complicated and diverse dataset to improve accuracy and generalisability was then done. Augmentation was done independently before each epoch with a randomly selected algorithm of the three augmentation algorithms. Our model took the augmented images (by one augmentation algorithm for each epoch; input) and calculated the probability of each image being a malignant diagnosis (output) after training a certain number of epochs (appendix p 6).

We used the weights of each network, pretrained on ImageNet, as the initialisation of our model's weights. The same training parameters were applied to each network branch. Stochastic gradient descent and cross-entropy loss were used for network weight tuning and algorithm optimisation. The initial learning rate was 0·01, which decreased by one-tenth every 100 epochs; the final learning rate was 0·0001. To prevent overfitting, batch normalisation was used and the weight decay rate was set to 0·0005. We used a batch size of 128 images and a Rectified Linear Unit activation function. Heatmaps were generated by the gradcam methods.

12 radiologists, including six junior radiologists (1–3 years of experience) and six senior radiologists (>8 years of experience), reviewed the two retrospective datasets and the prospective dataset. Radiologists were masked to the pathological confirmation of the nodule status and research aims before the reviewing process. The independent review process was made on a web-based rating platform. The review of each lesion included assigning points based on the ACR TI-RADS[15] categories (composition, echogenicity, shape, margin, and echogenic foci) and determining a malignant or benign diagnosis (appendix pp 17–24).

For more on the **rating platform** see http://thyroid.xiaobaishiji.com

ThyNet was tested in three stages (figure 1). First, the diagnostic performance of ThyNet was compared with radiologists (with test set A); second, improvement in the diagnostic performance of radiologists when assisted by ThyNet was evaluated (with test set B); and third, the application of ThyNet in actual clinical practice was investigated (with test set C).

For the first stage, ultrasound images from four independent hospitals were used to compare the performance of ThyNet with radiologists. Radiologists were invited to review the images and make diagnoses independently. A review process was made on a web-based rating platform, which integrated the data of all validation datasets. The review of each lesion included the following assigning points based on five ACR TI-RADS 18 categories (composition, echogenicity, shape, margin, and echogenic foci) and determining a malignant or benign diagnosis. All data were deidentified before transfer to the investigators, and the radiologists were also masked to the pathological reports. The radiologists were informed of their diagnostic performance compared with ThyNet before the deep-learning system was used to aid their diagnosis.

Radiologists in two hospitals used ThyNet to aid the diagnostic process. Initial independent review and diagnosis were made by radiologists alone. The radiologist diagnosis was compared with a reference diagnosis from ThyNet. If the two did not match, the radiologists could then choose to adhere to their diagnosis or adopt the diagnosis from ThyNet as the final diagnosis. Both the initial and final assisted diagnosis were recorded.

ThyNet was tested in a real-world clinical setting in three hospitals. Initial independent review and diagnosis were made by 12 radiologists reviewing static images and a second diagnosis was obtained based on dynamic videos of the nodule. The 12 radiologists were the same individuals that assessed the images in test sets A and B. The final diagnosis was made after the ThyNet-assisted reference diagnosis. The three independent diagnostic records of initial, second, and final diagnosis for each radiologist were recorded.

In clinical practice of thyroid nodule management, a crucial decision following ACR TI-RADS scoring is whether subsequent fine needle aspiration is indicated. According to ACR TI-RADS, nodules that score 2 points or less do not need fine needle aspiration, in which case the probability of being benign (negative predictive

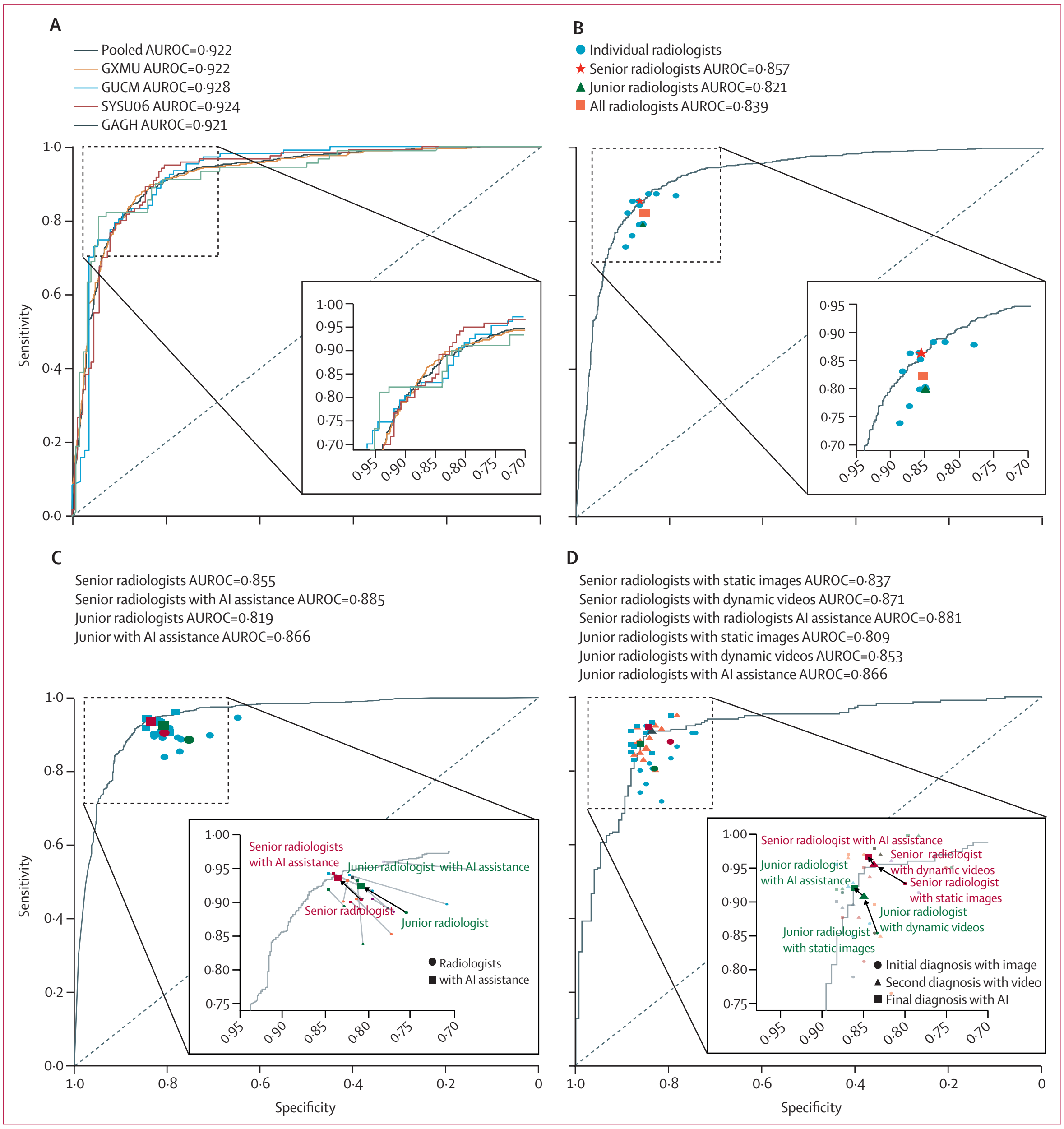

Figure 2: **Diagnostic performance of ThyNet and radiologists in serial test for discrimination of malignant from benign thyroid nodules**
(A) AUROCs to evaluate diagnostic performance of ThyNet in the total test set and each external institution in the first test comparing ThyNet with radiologists. (B) Diagnostic performance of ThyNet compared with each radiologist in the total test set. Round dots indicate diagnostic sensitivities and specificities of individual radiologists, the triangle indicates the pooled sensitivities and specificities of all junior radiologists, the star indicates the pooled sensitivities and specificities of all senior radiologists, and the square indicates pooled sensitivities and specificities of all radiologists. (C) Diagnostic performance of radiologists alone and radiologists assisted by ThyNet. Round dots indicate sensitivities and specificities of the first diagnosis, and the squares indicate sensitivities and specificities of second diagnosis with ThyNet assistance. (D) Diagnostic performance of radiologists assisted by ThyNet in a clinical setting. Round dots indicate sensitivities and specificities of the first diagnosis based on static images, triangles indicate the second diagnosis based on dynamic videos, and the squares indicate final diagnosis of radiologist with ThyNet assistance. AI=artificial intelligence. AUROC=area under the receiver operating characteristic curve. GAGH=the Guangzhou Army General Hospital. GUCM=the First Affiliated Hospital of Guangzhou University of Chinese Medicine. GXMU=the First Affiliated Hospital of Guangxi Medical University. ROC=receiver operating characteristic curve. SYSU06=the Sixth Affiliated Hospital of Sun Yat-sen University.

	AUROC (95% CI)	p value	Accuracy (95% CI)	p value	Sensitivity (95% CI)	p value	Specificity (95% CI)	p value	PPV	NPV	κ	F_1
ThyNet-only	0·944 (0·940–0·947)	..	0·891 (0·886–0·895)	..	0·949 (0·945–0·953)	..	0·812 (0·803–0·820)	..	0·871 (0·866–0·877)	0·922 (0·916–0·928)	0·772	0·909
Radiologists without ThyNet assistance												
All	0·837 (0·832–0·842)	<0·0001*	0·845 (0·840–0·850)	<0·0001*	0·894 (0·889–0·900)	<0·0001*	0·779 (0·771–0·788)	<0·0001*	0·845 (0·839–0·851)	0·846 (0·838–0·854)	0·680	0·869
Senior	0·855 (0·848–0·862)	<0·0001*	0·863 (0·856–0·869)	<0·0001*	0·904 (0·897–0·912)	<0·0001*	0·806 (0·795–0·818)	0·555*	0·863 (0·854–0·871)	0·862 (0·851–0·873)	0·716	0·883
Junior	0·819 (0·811–0·826)	<0·0001*	0·828 (0·821–0·836)	<0·0001*	0·885 (0·876–0·893)	<0·0001*	0·753 (0·740–0·765)	<0·0001*	0·828 (0·818–0·837)	0·829 (0·817–0·841)	0·644	0·856
Radiologists with ThyNet assistance												
All	0·875 (0·871–0·880)	<0·0001†	0·883 (0·879–0·888)	<0·0001†	0·929 (0·925–0·934)	<0·0001†	0·821 (0·813–0·829)	<0·0001†	0·875 (0·869–0·881)	0·896 (0·889–0·903)	0·758	0·901
Senior	0·885 (0·879–0·891)	<0·0001†	0·892 (0·886–0·898)	<0·0001†	0·935 (0·928–0·941)	<0·0001†	0·835 (0·824–0·846)	<0·0001†	0·884 (0·876–0·892)	0·905 (0·896–0·914)	0·777	0·909
Junior	0·866 (0·859–0·872)	<0·0001†	0·874 (0·868–0·880)	<0·0001†	0·924 (0·917–0·930)	<0·0001†	0·808 (0·796–0·819)	<0·0001†	0·866 (0·857–0·874)	0·887 (0·877–0·897)	0·739	0·894

AUROC=area under the receiver operating characteristic curve. F_1=a weighted average of the PPV and sensitivity. NPV=negative predictive value. PPV=positive predictive value. *Performance of ThyNet-assisted process with ThyNet alone. †Comparison of diagnostic performance without ThyNet.

***Table:* The diagnostic performance of ThyNet alone, radiologists alone, and ThyNet-assisted radiologists**

value) is high. Whereas, nodules with a score of 7 or more had a high probability of malignancy and a positive predictive value of pathological fine needle aspiration is reported to be between 76·6–95·4%.[16–18] If the ThyNet-assisted strategy can provide a similar positive predictive value as fine needle aspiration, the invasive procedure might not be warranted. Therefore, to alleviate the medical burden caused by growing need for fine needle aspirations, we simulated the scenario by omitting fine needle aspirations for nodules at a high probability of being benign (negative predictive value) or of high malignancy risk (positive predictive value), based on human ACR TI-RADS and ThyNet-assisted diagnosis.

We calculated the negative predictive value of nodules with ACR TI-RADS score 0–2. We also evaluated the negative predictive value of nodules with ACR TI-RADS scores of 3–6, diagnosed as benign with the ThyNet assistance process. We expected the score range could be expanded without a loss in negative predictive value compared with ACR TI-RADS scores of 0–2; therefore, fine needle aspirations could be omitted for nodules diagnosed as benign within this expanded score. We evaluated the positive predictive values of nodules diagnosed as malignant with the ThyNet-assistance process to acquire the score range with a positive predictive value more than 95%. For these nodules, we considered what additional medical intervention could be done without receiving fine needle aspirations. For all other nodules, medical recommendations were in accordance with the ACR TI-RADS guideline.

Statistical analysis

All statistical analyses were done using R (version 3.5.0). Comparisons with two-sided p values less than 0·05 were statistically significant.

ROC analyses with the ROCR package (version 1.15.0) were used to evaluate the diagnostic performance of ThyNet and radiologists when classifying malignant and benign lesions. For each test set, a ROC was created by plotting the true positive rate (sensitivity) against the true negative rate (specificity) by varying the predicted probability threshold. Subsequently, the AUROC values were calculated accordingly.

A two by two confusion matrix with the number of true positive, false positive, false negative, and true negative values was generated for each diagnosis. For thyroid cancer detection, the accuracy, sensitivity, specificity, positive predictive value, and negative predictive value were calculated according to the confusion matrix, and the 95% CIs for each value were calculated by the exact Clopper-Pearson method with GenBinomApps (version 1.0-2).[19]

The inter-radiologists agreement rate and Fleiss' κ value to evaluate the agreement of the diagnosis of malignant and benign lesions were calculated for each test set by the kappam.fleiss function, R software (irr version 0.84.1). The intraclass correlation coefficient to evaluate the agreement between the 12 radiologts regarding the five ACR grades were calculated by using the icc function of the irr package (version 0.84.1).[20] The fine needle aspiration rate and missed malignancy rate between ACR TI-RADS with or without ThyNet-assisted fine needle aspiration strategy were compared using McNemar's χ^2 test.

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

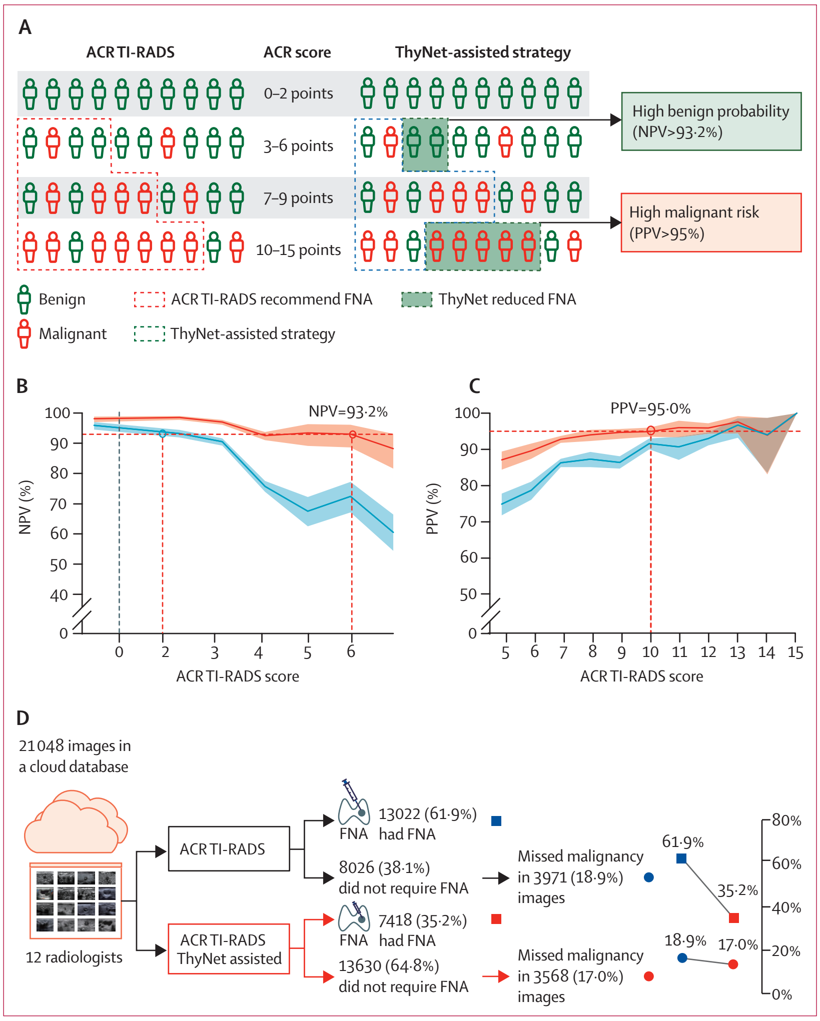

Figure 3: **Recommendation for FNA with ThyNet assistance**

(A) FNA strategy with ThyNet assistance. With the assistance of ThyNet, an ACR TI-RADS score of 3–6 points for some nodules can achieve a similarly high NPV of 2 points, which is higher than or equal to 93·2%. Whereas an ACR TI-RADS score of 10–15 points for some nodules can achieve a similarly high PPV to FNA, which is higher than or equal to 95·0%. According to the ACR TI-RADS, no FNA was needed if nodules were scored 2 points, whose NPV for being benign was 93·2%. With the assistance of ThyNet, an ACR TI-RADS score of 3–6 points for some nodules can achieve a similarly high NPV of 2 points. Whereas an ACR TI-RADS score of 10–15 points for some nodules can achieve a similarly high PPV to FNA, whose PPV reached up to 95%. (B) Trend of NPV according to ACR scores when diagnosis is benign. The shaded areas are 95% CI. (C) Trend of PPV according to ACR scores when diagnosis is malignant. The shaded areas are 95% CI. (D) Thyroid nodule management assisted by ThyNet. The ThyNet-assisted strategy reduced FNA from 61·9% to 35·2%, while missed malignancy decreased from 18·9% to 17·0%. ACR=American College of Rheumatology. FNA=fine needle aspiration. NPV=negative predictive value. PPV=positive predictive value. TI-RADS=Thyroid Imaging Reporting and Data System.

Results

ThyNet for differentiating malignant tumours from benign nodules in screening ultrasound was developed and trained using images from the First Affiliated Hospital of Sun Yat-sen datasets University and Sun Yat-sen University Cancer Center (training dataset 14439 images of 6587 patients and the tuning dataset 3610 images of 1752 patients).

2185 images of 1424 patients with thyroid nodules—enrolled from four independent hospitals (test set A)—were used to test ThyNet against diagnosis made by radiologists. 1754 images of 1048 patients with thyroid nodules—enrolled from two hospitals (test set B)—were used to test the ThyNet-aided diagnostic process. 366 images of 303 patients with thyroid nodules—enrolled from three hospitals (test set C)—were used to test ThyNet in a real-world clinical setting. The demographic characteristics of patients in retrospective and prospective datasets are shown in the appendix (p 9). Papillary thyroid cancer was the most prevalent thyroid cancer subtype in training and total test sets (appendix p 10).

In test set A, the AUROC of ThyNet-only (0·922 [95% CI 0·910–0·934]) was significantly higher than that of the radiologists-only (0·839 [0·834–0·844]; p<0·0001; figure 2A, B). The accuracy of ThyNet (0·861 [95% CI 0·845–0·876]) was also significantly higher than that of the radiologists (0·841 [0·83–0·846]; p<0·0001); the AUROC was 0·857 (0·850–0·863) for senior radiologists and 0·821 (0·814–0·828) for junior radiologists (appendix p 11). Basic information about the radiologists is provided in the appendix (p 12). Heatmaps of ThyNet analysing nodules are in the appendix (p 12).

In the assessment of the ThyNet-aided diagnostic process (test set B; appendix p 13), the pooled AUROC for the radiologists alone was 0·837 (95% CI 0·832–0·842), which was improved to 0·875 (0·871–0·880) with ThyNet assistance (p<0·0001; table and figure 2C). The pooled accuracy for the radiologists alone was 0·845 (95% CI 0·840–0·850), which was improved to 0·883 (0·879–0·888) with ThyNet assistance (p<0·0001). For the senior radiologists, the pooled AUROC improved from 0·855 (0·848–0·862) to 0·885 (0·879–0·891; p<0·0001) and for the junior radiologists the pooled AUROC improved from 0·819 (0·811–0·826) to 0·866 (0·859–0·872; p<0·0001) with ThyNet assistance (table). With the assistance of ThyNet, the accuracy of the junior radiologists improved from 82·8% to 87·4% (p<0·0001), and the κ value increased from 0·644 to 0·739 (p<0·0001).

In the real-world clinical setting test of ThyNet (test set C), the pooled AUROC of the initial diagnosis (radiologists reviewing static images only) was 0·823 (95% CI 0·812–0·835); the AUROC of the second diagnosis (radiologists reviewing videos and images) was improved to 0·862 (0·851–0·872; p<0·0001); in the final diagnosis with ThyNet assistance, the AUROC was improved to 0·873 (0·863–0·883; p<0·0001; figure 2D). The pooled accuracy of the initial diagnosis was 0·824 (95% CI 0·813–0·836); the accuracy of the second diagnosis was improved to 0·865 (0·855–0·875; p<0·0001); in the final diagnosis with ThyNet assistance, the accuracy was improved to 0·877 (0·867–0·886, p=0·1193; appendix p 14). Of note, the diagnostic performance was improved from image to video to ThyNet assistance, for both senior (p=0·0075) and junior (p<0·0001) radiologists (appendix p 14). In a post-hoc analysis, There was no significant difference in the diagnostic accuracy of

ThyNet between papillary thyroid cancer and follicular thyroid cancer (appendix p 15).

The classifications made by the ThyNet system could be used to reduce the workload involved in the decision-making process of fine needle aspiration, while preserving the standard of care. We simulated the scenario by omitting fine needle aspirations at high negative predictive value and positive predictive value when ACR TI-RADS accommodated the ThyNet decision (figure 3A).

According to the ACR TI-RADS guideline, no fine needle aspiration was needed if nodules had a score of 2 points or less. The negative predictive value for predicting benign nodules for these nodules reached 93·2% (95% CI 91·9–94·4). Score-related negative predictive values were recalculated when the ThyNet diagnosis was also benign. When ACR TI-RADS was used with the assistance of ThyNet, the negative predictive values of thyroid nodules scored 3 were as high as 97%; 92·5% for nodules scored 4, 93·2% for nodules scored 5, and 93·0% for nodules scored 6 (figure 3B). The size distribution of the nodules with ACR TI-RADS scores of 3–6 are listed in the appendix (p 16). Of the 1754 nodules, a median of 315 (IQR 250–337) procedures could be avoided with the assistance of Thynet due to the high probability of being benign.

The positive predictive value of pathological fine needle aspiration for thyroid nodules is between 76·6–95·4%.[16–18] With the assistance of ThyNet, nodules diagnosed as malignant and scored from 10 to 15 points reached a similar positive predictive value at 95% (figure 3C). Of the 1754 nodules assessed, a median of 134 (IQR 109–199) fine needle aspirations could have been avoided with the assistance of ThyNet due to the high probability of being malignant.

Of the 21 048 diagnoses made by 12 radiologists, the ThyNet-assisted strategy reduced fine needle aspirations from 61·9% to 35·2% (p<0·0001), while missed malignancy decreased from 18·9% to 17·0% (p<0·0001; figure 3D).

Discussion

To our knowledge, this study is the first to evaluate a deep-learning strategy as an aid for thyroid nodule management. The ThyNet-assisted strategy not only improved the diagnostic accuracy of radiologists when reviewing images only but also when reviewing images and videos in a clinical setting. Of note, the combination of the ACR TI-RADS classification with AI assistance improved the negative predictive value and positive predictive value, which has the potential to reduce the number of unnecessary fine needle aspirations.

This study focused on developing a deep-learning AI-assisted strategy for clinical decision making regarding thyroid nodules. Previous studies have indicated that deep-learning algorithms outperformed health-care professionals with respect to some clinical outcomes.[3,12,21] However, the stand-alone application of an AI diagnostic model is not practical for clinical use. In medical workflows, noise data, including mismatches, omissions, and errors, go far beyond the scope of these ideal algorithms. Real-world decisions should be supervised by clinicians even if AI is reported to have superior performance. Therefore, the most important role of this AI model is improving diagnostic accuracy by assisting clinicians in treatment decisions.

On the basis of our findings, we make two recommendations for the implementation of AI systems into clinical practice. First, our model showed advantages in improving the accuracy of diagnosis, especially for junior radiologists. The interpretation of ultrasound images is often subject to significant interobserver variabilities, particularly for junior radiologists at non-academic centres, with reported variability in sensitivity from 40·3% to 100% and in specificity from 50% to 100% radiologists.[22–25] ThyNet would provide a consistent second opinion on nodule images. Results showed that junior radiologists tend to accept ThyNet's advice more often than senior radiologists, which might be related to less confidence and less working experience. With the assistance of ThyNet, the accuracy of the junior radiologists improved from 82·8% to 87·4%. In the clinical setting test, sonographic assessment of thyroid nodules included both real-time dynamic nodule visualisation and interpretation of static images. ThyNet also improved the performance of radiologists in this clinical setting. Second, the ThyNet-assisted fine needle aspiration strategy could be useful for the avoidance of unnecessary invasive biopsy. The latest ACR TI-RADS system, used in this study, was reported to have the lowest unnecessary fine needle aspiration rates in published guidelines.[26,27] The nodules scored 2 points or less do not need fine needle aspiration,[15] of which the negative predictive value was 93·2% in our study. However, the range could be expanded to 6 points without a clinically significant loss in negative predictive value (93·0% at 6 points) if applied with our strategy. For suspected malignant nodules, all guidelines recommend fine needle aspiration before surgery. However, approximately 15–30% of fine needle aspirations will be assigned to uncertain cytopathology categories.[28,29] The reported positive predictive value of fine needle aspiration is approximately 76·6–95·4%.[16–18] Therefore, because the positive predictive value in this study was more than 95%, recommending surgery directly without fine needle aspiration could be possible. With the assistance of ThyNet, an ACR TI-RADS score from 10 to 15 achieved a positive predictive value of more than 95%, and the ThyNet-assisted fine needle aspiration strategy also provided a modest advantage in reducing missed malignancy. Fine needle aspiration is not recommended for small nodules (<1 cm in diameter) with an ACR TI-RADS score of 10–15 because of the high false negative rate associated with the process.[15] However because this is a simulation, in the ThyNet-assisted strategy, these nodules with a high malignancy risk would be recommended for additional medical intervention. Overall, assistance with ThyNet might reduce the number of fine needle aspirations by 26·7% and the number of missed malignancies by 1·9%. Of note, some malignancies were still missed because nodules with a diameter between 3 mm and 1 cm were not recommended for fine needle aspiration. The percentage of missed malignancies in our study is not substantially different from what has been reported for missed under ACR TI-RADS.[30]

This study has several limitations. All patients in the training set and the total test set underwent pathological examination in a hospital setting instead of a screening setting for thyroid nodules. The diverse prevalence might significantly affect the positive predictive value and negative predictive value between populations, which could decrease the potential generalisability of the results. The accuracy and stability of the deep-learning model relies on gold standards of cytopathological and surgical pathology diagnosis, which are hard to obtain in the general population in a real-world setting. The main objective of this study was to evaluate the improvement in accuracy of diagnosis with ThyNet assistance to potentially reduce unnecessary fine needle aspiration procedures in the hospital setting. Therefore, interpretation of these results might have its own confines that cannot be discreetly extrapolated to other contexts. Another potential limitation is that in the first test of ThyNet as a diagnostic aid, the diagnoses were made by reviewing ultrasound images only and so the performance of radiologists could be underestimated. However, in the clinical setting, ThyNet still improved the performance of radiologists after reviewing ultrasound images and video. Compared with a previous study,[12] a better performance of radiologists alone was observed in this study. Moreover, potential biases exist with respect to the selection of radiologists and data, including the exclusion of low-quality images and the exclusion of normal scans. And for patients with multiple nodules, challenges exist with nodule-path correlation in the retrospective dataset. Therefore, future validation in a large-scale screening cohort is needed.

In conclusion, the ThyNet-assisted strategy significantly improved the diagnostic accuracy of radiologists on thyroid nodule differentiation and could potentially decrease the number of unnecessary fine needle aspirations.

Contributors

HaiX, WW, and EKA supervised the study. HaiX, SP, WW and YiL conceived of and designed the study. GH and YG trained and developed the artificial intelligence model. QiaZ did the statistical analysis. HaiX, SP, WW, YiL, QiaZ, HH, HanX, and ZL wrote the drafted report. HW and EKA critically revised the manuscript. YaL, JinL, JieL, FL, QiuZ, QD, FN, HY, JR, GL, XW, and XZ organised and screened patients. All authors had access to all the raw datasets. HaiX and SP verified all the data. All authors revised the report and approved the final version before submission.

Declaration of interests

We declare no competing interests.

Data sharing

The source codes used in this study are available online. Authors will share deidentified individual participant imaging data on request with researchers who provide a methodologically viable proposal and can do analyses that achieve the aims of the proposal. Data sharing requests can be directed to XiaX by email. To gain access data requestors will need to sign a data access agreement.

For more on the **source codes** see https://github.com/sprint2200/ThyNet

Acknowledgments

This study was funded by grants from the National Natural Science Foundation of China (81772850 and 81971630), and the Guangzhou Science and Technology Project (201803010057 and 201904010187).

References

1 Guth S, Theune U, Aberle J, Galach A, Bamberger CM. Very high prevalence of thyroid nodules detected by high frequency (13 MHz) ultrasound examination. *Eur J Clin Invest* 2009; **39:** 699–706.

2 Haugen BR, Alexander EK, Bible KC, et al. 2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: the American Thyroid Association Guidelines Task Force on thyroid nodules and differentiated thyroid cancer. *Thyroid* 2016; **26:** 01–133.

3 Choi SH, Kim EK, Kwak JY, Kim MJ, Son EJ. Interobserver and intraobserver variations in ultrasound assessment of thyroid nodules. *Thyroid* 2010; **20:** 167–72.

4 Alexander EK, Kennedy GC, Baloch ZW, et al. Preoperative diagnosis of benign thyroid nodules with indeterminate cytology. *N Engl J Med* 2012; **367:** 705–15.

5 Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; **542:** 115–18.

6 Ting D, Cheung CY, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA* 2017; **318:** 2211–23.

7 Liang J, Huang X, Hu H, et al. Predicting malignancy in thyroid nodules: radiomics score versus 2017 american college of radiology thyroid imaging, reporting and data system. *Thyroid* 2018; **28:** 1024–33.

8 Thomas J, Haertling T. AIBx, Artificial intelligence model to risk stratify thyroid nodules. *Thyroid* 2020; **30:** 878–84.

9 De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med* 2018; **24:** 1342–50.

10 Kermany DS, Goldbaum M, Cai W, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* 2018; **172:** 1122–31.e9.

11 Buda M, Wildman-Tobriner B, Hoang JK, et al. Management of thyroid nodules seen on US images: deep learning may match performance of radiologists. *Radiology* 2019; **292:** 695–701.

12 Li X, Zhang S, Zhang Q, et al. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. *Lancet Oncol* 2019; **20:** 193–201.

13 Wei X, Gao M, Yu R, et al. Ensemble deep learning model for multicenter classification of thyroid nodules on ultrasound images. *Med Sci Monit* 2020; **26:** e926096.

14 Huang G, Liu Z, Maaten LVD, Weinberger KQ. Densely connected convolutional networks. IEEE Conference on Computer Vision and Pattern Recognition; Honolulu, HI, USA; 21–26 July 2017.

15 Tessler FN, Middleton WD, Grant EG, et al. ACR thyroid imaging, reporting and data system (TI-RADS): white paper of the ACR TI-RADS committee. *J Am Coll Radiol* 2017; **14:** 587–95.

16 Lee J, Lee SY, Cha SH, Cho BS, Kang MH, Lee OJ. Fine-needle aspiration of thyroid nodules with macrocalcification. *Thyroid* 2013; **23:** 1106–12.

17 Yoon JH, Kwak JY, Moon HJ, Kim MJ, Kim EK. The diagnostic accuracy of ultrasound-guided fine-needle aspiration biopsy and the sonographic differences between benign and malignant thyroid nodules 3cm or larger. *Thyroid* 2011; **21:** 993–1000.

18 Lee TI, Yang HJ, Lin SY, et al. The accuracy of fine-needle aspiration biopsy and frozen section in patients with thyroid cancer. *Thyroid* 2002; **12:** 619–26.

19 Clopper C, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 1934, **26:** 404–13.

20 Leijenaar RT, Carvalho S, Velazquez ER, et al. Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability. *Acta Oncol* 2013; **52:** 1391–97.

21 Ko SY, Lee JH, Yoon JH, et al. Deep convolutional neural network for the diagnosis of thyroid nodules on ultrasound. *Head Neck* 2019; **41:** 885–91.

22 Kim HG, Kwak JY, Kim EK, Choi SH, Moon HJ. Man to man training: can it help improve the diagnostic performances and interobserver variabilities of thyroid ultrasonography in residents? *Eur J Radiol* 2012; **81:** e352–56.

23 Park CS, Kim SH, Jung SL, et al. Observer variability in the sonographic evaluation of thyroid nodules. *J Clin Ultrasound* 2010; **38:** 287–93.

24 Park SH, Kim SJ, Kim EK, Kim MJ, Son EJ, Kwak JY. Interobserver agreement in assessing the sonographic and elastographic features of malignant thyroid nodules. *AJR Am J Roentgenol* 2009; **193:** W416–23.

25 Kim SH, Park CS, Jung SL, et al. Observer variability and the performance between faculties and residents: US criteria for benign and malignant thyroid nodules. *Korean J Radiol* 2010; **11:** 149–55.

26 Xu T, Wu Y, Wu RX, et al. Validation and comparison of three newly-released thyroid imaging reporting and data systems for cancer risk determination. *Endocrine* 2019; **64:** 299–307.

27 Ha EJ, Na DG, Moon WJ, Lee YH, Choi N. Diagnostic performance of ultrasound-based risk-stratification systems for thyroid nodules: comparison of the 2015 American Thyroid Association guidelines with the 2016 Korean Thyroid Association/Korean Society of Thyroid Radiology and 2017 American Congress of Radiology guidelines. *Thyroid* 2018; **28:** 1532–37.

28 Nikiforov YE, Carty SE, Chiosea SI, et al. Highly accurate diagnosis of cancer in thyroid nodules with follicular neoplasm/suspicious for a follicular neoplasm cytology by ThyroSeq v2 next-generation sequencing assay. *Cancer* 2014; **120:** 3627–34.

29 Suh YJ, Son EJ, Moon HJ, Kim EK, Han KH, Kwak JY. Utility of thyroglobulin measurements in fine-needle aspirates of space occupying lesions in the thyroid bed after thyroid cancer operations. *Thyroid* 2013; **23:** 280–88.

30 Magri F, Chytiris S, Croce L, et al. Performance of the ACR TI-RADS and EU TI-RADS scoring systems in the diagnostic work-up of thyroid nodules in a real-life series using histology as reference standard. *Eur J Endocrinol* 2020; **183:** 521–28.